

ESD-TR-65-542

ESD ACCESSION LIST

ESTI Call No. AI 50682Copy No. 1 of 1 cys

---

**ESD RECORD COPY**

RETURN TO  
SCIENTIFIC & TECHNICAL INFORMATION DIVISION  
(ESTI), BUILDING 1211

*100-100* DIRECT VS INDIRECT ASSESSMENT OF  
SIMPLE KNOWLEDGE STRUCTURES

H. Edward Massengill  
Emir H. Shuford, Jr.

March 1966

DECISION SCIENCES LABORATORY  
ELECTRONIC SYSTEMS DIVISION  
AIR FORCE SYSTEMS COMMAND  
UNITED STATES AIR FORCE  
L. G. Hanscom Field, Bedford, Massachusetts

Distribution of this document  
is unlimited.



ESRH

A00632609

When US Government drawings, specifications or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Do not return this copy. Retain or destroy.

DIRECT VS INDIRECT ASSESSMENT OF  
SIMPLE KNOWLEDGE STRUCTURES

H. Edward Massengill  
Emir H. Shuford, Jr.

March 1966

DECISION SCIENCES LABORATORY  
ELECTRONIC SYSTEMS DIVISION  
AIR FORCE SYSTEMS COMMAND  
UNITED STATES AIR FORCE  
L. G. Hanscom Field, Bedford, Massachusetts

Distribution of this document  
is unlimited.

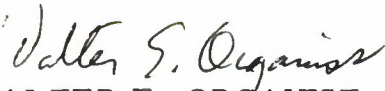


## FOREWORD

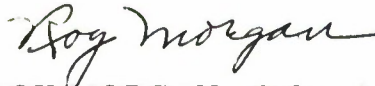
This study was conducted in support of Project 2806, Task 280609, over the period May 1964 through January 1965.

This report presents a formal comparison between two types of classroom testing (a) traditional multiple choice and (b) distribution of uncertainty over choices.

This Technical Report has been reviewed and is approved.



WALTER E. ORGANIST  
Project Scientist  
Decision Sciences Laboratory



ROY MORGAN, Colonel, USAF  
Director  
Decision Sciences Laboratory



## ABSTRACT

This report compares two types of classroom testing in terms of efficacy in guiding instruction. One type of testing is the traditional indirect method based on the observation of choices. The other type is the direct method based on admissible probability measurement. The general finding is that the direct methods always perform as well as and in most cases better than the indirect methods. This deficiency in the indirect method can be alleviated in theory by introducing redundancy into the test and asking the same question over and over again. The performance of indirect methods depends in a very critical manner upon the information available to the instructor from other sources about the current state of knowledge of each student. The performance of the direct methods is unaffected by this. The gain in effectiveness achieved by using direct methods must be balanced off against the cost of using these new methods. A direct method may require more student time per item than does an indirect method. This, however, may be more than compensated for by the requirement for redundancy when using the indirect method. In addition, since a direct method does not require additional information from the instructor as to the current state of knowledge of each student, the possibility exists that much larger classes may be taught with no loss in effectiveness thus implying even further economic benefits from the use of direct methods to guide classroom instruction.

## TABLE OF CONTENTS

	PAGE
1. Statement of the Problem	1
2. Classification Without Testing	4
3. The Response Model for the Two Methods of Testing	8
4. Formal Analysis of the Two Testing Methods	14
4.1 Direct Method	15
4.2 Indirect Method	18
5. The Effectiveness of the Two Methods	30
5.1 Effectiveness from the Instructor's Point of View	31
5.2 Effectiveness from the Outside Agent's Point of View	37
5.3 Summary	45
References	47

# DIRECT VS. INDIRECT ASSESSMENT OF SIMPLE KNOWLEDGE STRUCTURES

H. Edward Massengill and Emir H. Shuford, Jr.

## 1. Statement of the Problem.

In this report we compare, mathematically, two testing methods in a well-defined situation. Our purpose is to determine how the two methods perform in the matter of classifying students in this situation and to ascertain some of the distinguishing characteristics of each method. Since our results are logically derived from explicitly stated assumptions, we have no doubt as to their validity for the specific situation we are examining. Further, if there are real-life situations which are equivalent to the one we define, we can be certain that our results will apply to these situations. But we will not be concerned here in seeking to determine the extent of the generality of the situation we have chosen. This is not crucial for our purpose. This should not be taken to mean that we are not concerned with how these results may relate to more complex situations. On the contrary, we hope that the findings for this situation will give us a better idea of what to look for in more complex situations. And we are confident that the approach we have used, namely the application of purposive mathematics (Massengill, 1964), can be extended to aid us in the analysis of these more complex situations.

The two methods which we will compare are the traditional indirect method, IM\*, and the direct method, DM. In the indirect method, the

---

\*

We intend to deal with the indirect method in terms of decision theory so that all of the information available to the person using this method may be explicitly taken into account.

student is given a question with two or more alternatives and asked to choose the correct alternative. In the direct method, the student is also given a question with two or more alternatives. But instead of being asked to give the correct answer, he interacts with a measurement procedure which outputs an inferred subjective probability distribution over the alternatives.\*

In order for the results of our comparison to be meaningful, we must know exactly what assumptions are involved both in the student's response process and in the two testing methods. To keep the assumptions simple, and thereby make the arguments easier to follow, we will use a very simple, but not unrealistic, testing situation. The test will consist of one two-alternative question, or the same two-alternative question repeated several times. The purpose of the test will be to help determine if a student knows a particular concept.

The concept which we will test deals with the classification of two objects,  $B$  and  $C$ , according to whether  $A=B$  or  $A=C$ . A given student has been through a lesson in which the instructor has taught that  $A=B$  and that it is not the case that  $A=C$ . If the student has learned  $A=B$ , we shall say that he is trained,  $T$ . If he has learned that  $A=C$ , we shall say that he is mis-trained,  $mT$ . If he has learned nothing, we shall say that he is untrained,  $uT$ .

At the close of the lesson, we want to classify the student as belonging to one of the three categories:  $T$ ,  $uT$ , or  $mT$ . The student's next lesson

---

\*

See Shuford (1965), Shuford, Albert, & Massengill (1965), and Shuford & Massengill (1965) for more information about these measurement procedures.



will depend on how he is classified at the end of this lesson. Thus, if he is classified as trained, he will go on to the next lesson. If he is classified as untrained, the same lesson will be repeated. And if he is classified as mistrained, he will be given the same lesson in a different form. Because of the effect of the classification on the next step of the student's training, we will use a payoff scheme for which a correct classification is more valuable than an incorrect one and for which the values of correct classifications are equal and the values of incorrect classifications are equal. For the derivations of this report, a correct classification will have a value of 1.0 and an incorrect one a value of 0. Table 1 shows the payoff matrix for this decision problem.

TABLE 1

Payoff Matrix for the Instructor who is Classifying  
Students According to Three Categories

ACTS	CATEGORY			$EU(a_i)$
	$T$	$uT$	$mT$	
$a_1: T$	1.0	0	0	$P(T)$
$a_2: uT$	0	1.0	0	$P(uT)$
$a_3: mT$	0	0	1.0	$P(mT)$

It should be noted that the particular utility structure we are using makes the expected utility of an act equivalent to the expected proportion of correct classifications. Thus, each statement which we make about expected utility can also be interpreted in terms of expected proportion of correct classifications. For the most part, we will use the term expected utility since it is more general.

## 2. Classification Without Testing.

It is possible for the instructor to classify students with some accuracy even without giving them a formal test. After all, he may have observed the students during the lesson and may have some rather strong feelings, at least for some of the students, about which category a student is in. Suppose, for example, that a student spent the whole period of the lesson working on some other assignment. Then the instructor might have reason to believe that the student is untrained. Suppose that another student dozed during the lesson and only came to life during the time the instructor was discussing  $A=C$ . But suppose that he did not remain awake long enough to hear the instructor make the point that  $A=C$  is *not* the case. If the instructor noticed this, then he might be pretty sure that the student should be classified as mistrained. Finally, suppose that a third student had listened intently to the instructor's words during the lesson. Then the instructor might be fairly sure that this student was trained, since this concept should be easy to learn if one hears it explained. Thus, by observing the students during the lesson, the instructor might be able to do a fairly good job of classification without testing the students.

If an instructor can evaluate his subjective probabilities concerning which category a student is in and express them, then his expected utility can be determined for each combination of prior probability values.

Figure 1 shows the surface of all the possible combinations of the instructor's prior probability values for the three categories. The surface is divided into three sections each of which is characterized by

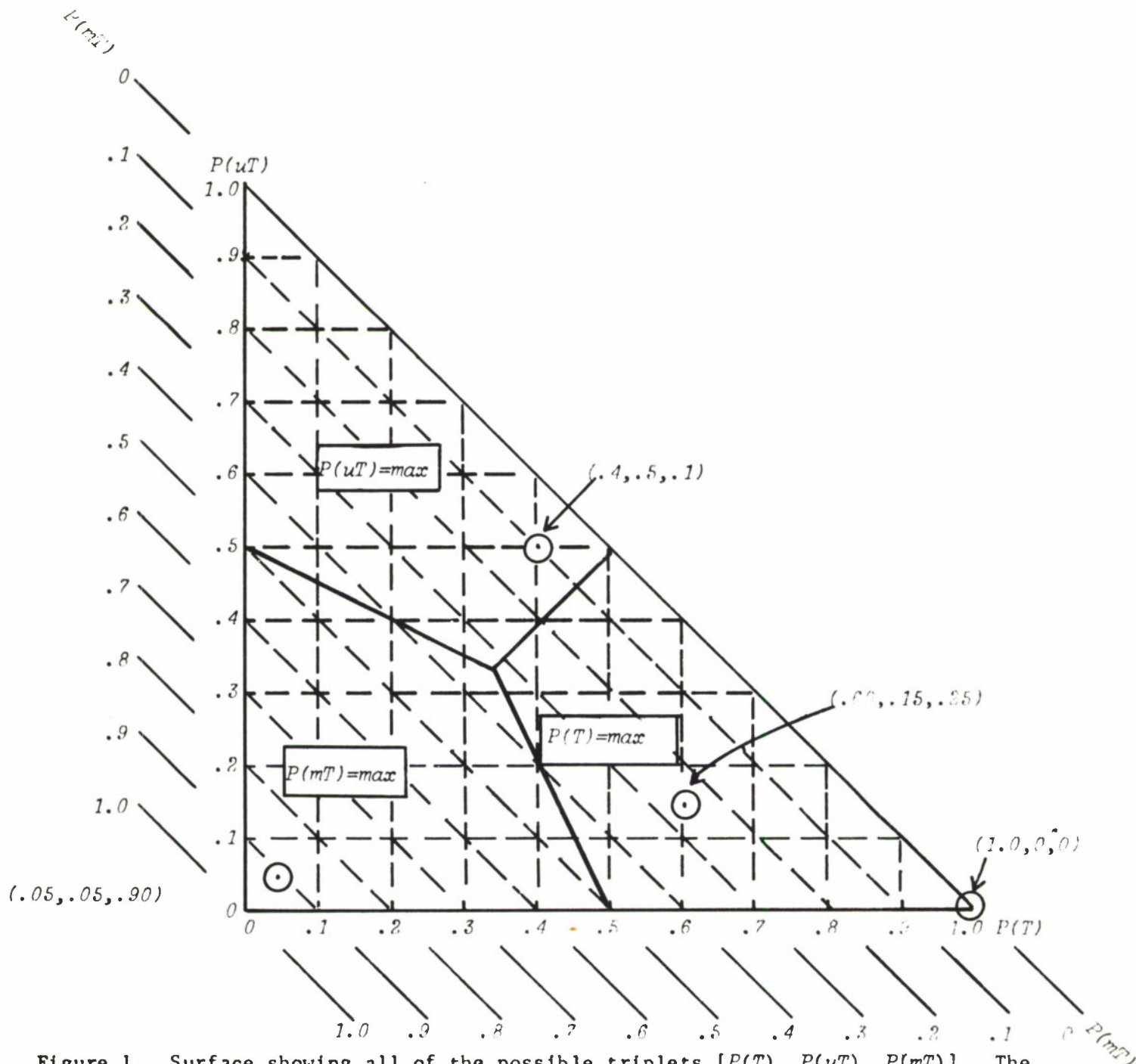


Figure 1. Surface showing all of the possible triplets  $[P(T), P(uT), P(mT)]$ . The value of any member of a triplet is  $\geq 0$  and  $P(T) + P(uT) + P(mT) = 1.0$ . The surface is divided into three areas. In each area a different one of the probabilities is a maximum for all points within that area.  $P(mT)$  is constant along a given slant line.

the fact that one of the three categories has the maximum probability for each point in that section. From Table 1, we see that the expected utility of a given act is equal to one of these three probabilities. The act which specifies the choice of the most probable category is the one which maximizes expected utility and the probability of the most probable category is equivalent to the maximum expected utility. Using this information, we can determine the expected utility associated with each possible prior probability combination.

Figure 2 represents this information in two dimensions. It is helpful to conceptualize the information contained in Figure 2 in terms of three dimensions. The base of the three-dimensional figure would be identical to Figure 1. The third axis of the figure would contain  $EU(a^*)$ , which is shown by the dashed lines in Figure 2.

Note that the minimum expected utility is  $1/3$  and that this occurs only for the probability combination  $(1/3, 1/3, 1/3)$ , i.e.,  $P(T)=P(uT)=P(mT)=1/3$ . The maximum expected utility,  $EU(a^*) = 1.0$ , occurs at three points, the three corners of the base of the figure. The values for the other points on the surface grow larger as they increase in distance from the point  $(1/3, 1/3, 1/3)$ .

If the instructor wants to decide whether or not to test and which kind of test to use, obviously, he will need to know how the expected utility of testing for each of the two methods compares with the expected utility of classifying without testing. Thus, we will need to obtain, for the two test methods, the information analogous to that which we obtained and summarized in Figure 2 for the decision to classify without testing.

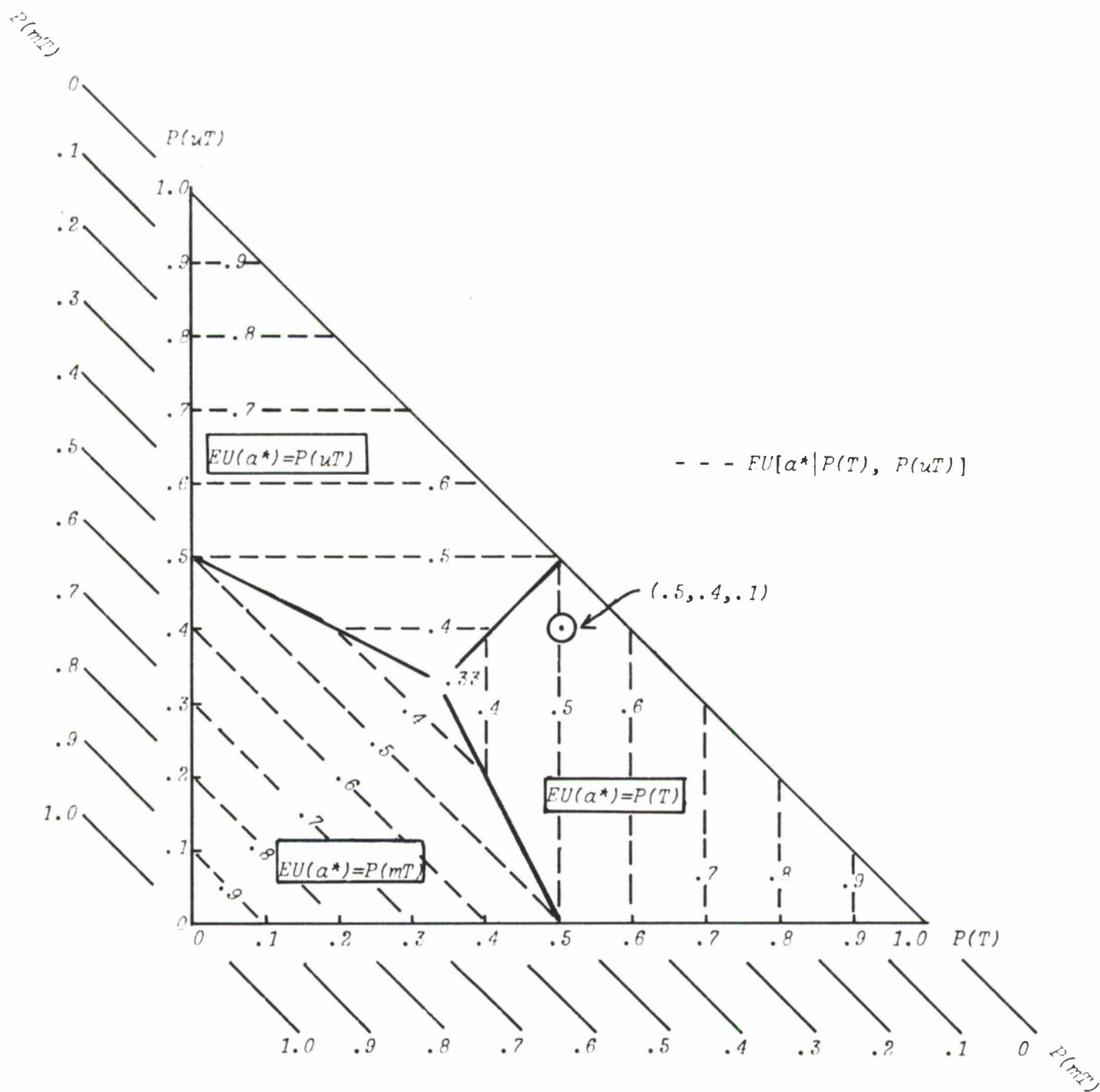


Figure 2. The dashed lines show the expected utility for all the points through which the lines pass. For example,  $EU(\alpha^*)$  for the point  $(.5, .4, .1)$  is .5.

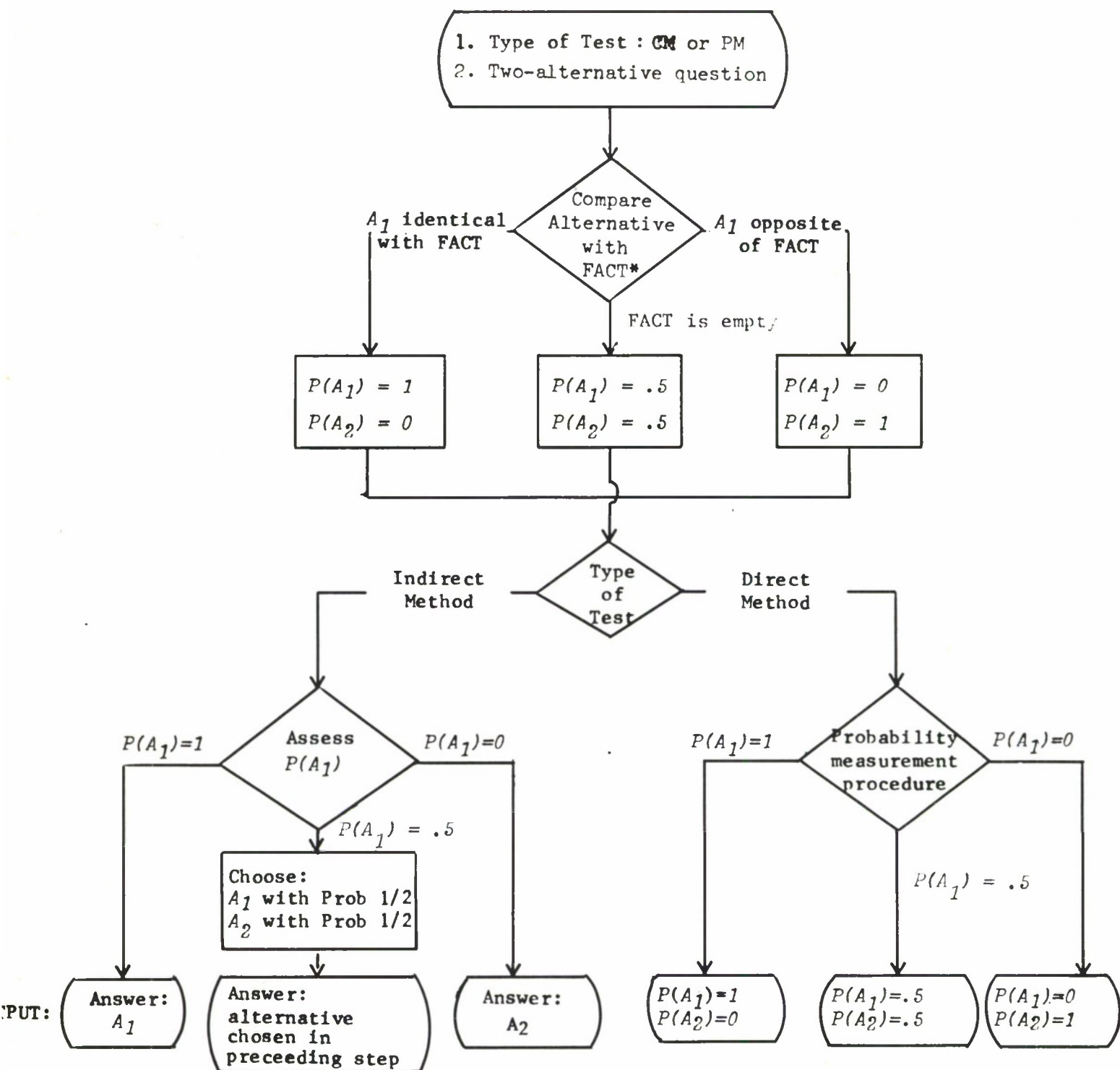


### 3. The Response Model for the Two Methods of Testing.

To begin with we will examine the model which generates the responses for the situation we are considering. In order to make our assumptions concerning the model explicit, we will describe it in terms of a task being performed by a machine. In applying our results to any real-life situation, the main concern will be to determine if the model is a good description of the behavior of the students in question. It might be helpful to point out once more that our main purpose is not to determine which, if any, actual situations the model describes but rather to study the performance of the two test methods for the model in an attempt to make statements about the value of each method and to get some idea of the important characteristics of each method.

Figure 3 shows the response model in the form of a computer flow chart. We will examine the logic of the flow chart step-by-step in order to make the assumptions in our response model as clear as possible. First, let us examine the location FACT. During the training period, one of three things happens to FACT. It takes the value  $A=B$ , corresponding to the category,  $T$ ; it takes the value  $A=C$ , corresponding to the category,  $mT$ ; or it remains empty, corresponding to the category,  $uT$ .

When the test is given, there are two pieces of information given as input to the machine: the type of testing method, indirect or direct, and the two-alternative question on the concept. The first step for the machine is to compare the first alternative,  $A_1$ , with FACT. If  $A_1$  and FACT are identical, then alternative 1 is given a probability of one, i.e.  $P(A_1)=1.0$ , and alternative 2 is given a probability of zero. If  $A_1$  is the opposite of



\*FACT contains:  
 A=B if trained  
 A=C if mistrained  
 Empty if untrained.

Figure 3. Flow chart of the model producing the responses for the two test methods.

FACT, then the probabilities of the two alternatives are the reverse of the case above. If FACT is empty, then each alternative is given a probability of .5.

Table 2 shows the two possible forms of our question, i.e., one with  $A=B$

TABLE 2

$P(A_i)$  Associated with Alternative  $i$ , where  $i=1$  or  $2$ , for a Given State of Knowledge and a Given Form of the Question

FORMS		CATEGORIES		
		TRAINED	UNTRAINED	MISTRAINED
$F_1$	$A_1: A=B$	1.0	.5	0
	$A_2: A=C$	0	.5	1.0
$F_2$	$A_1: A=C$	0	.5	1.0
	$A_2: A=B$	1.0	.5	0

as the first alternative and the other with  $A=C$  as the first alternative. Of course, the question could also be asked in true-false form. The table also shows the three possible states and the results that our model would yield for  $P(A_i)$  for a given form of the question.

It is immediately evident from Table 2 that if we could obtain  $P(A_i)$  from a given student, we could correctly classify him with one question, for this situation. For example, if we gave a student the first form of question,  $F_1$  and found that  $P(A_1)=1.0$ , we would know that he was trained.

If we found that  $P(A_1)=.5$ , we would know that he was untrained. And if we found that  $P(A_1)=0$ , we would know that he was mistrained. Getting  $P(A_i)$  is exactly the purpose of DM. Thus, for this situation, DM would give us perfect classification with a one-item test.

After comparing  $A_1$  with FACT, the machine then determines which testing method is being used and acts accordingly. If the DM is being used, the machine goes to a probability measurement procedure and it is this measurement procedure which actually outputs the probabilities for the two alternatives, after having inferred them from the results of an interrogation of the machine.

If IM is being used, then the machine chooses one of the alternatives as being correct. The machine makes its choice in terms of  $P(A_1)$ . This is not an arbitrary procedure but is based on the machine's payoff matrix. The payoff matrix is shown in Table 3. The utility structure of the payoff matrix is "all-or-none", i.e., a correct outcome is more valuable than an incorrect one, the correct outcomes all have equal values, and the incorrect outcomes all have equal values. If we represent the value of a correct outcome with 1.0 and the value of an incorrect outcome with 0,

TABLE 3

The Subject's Payoff Matrix for the Indirect Method of Testing

ACTS	CATEGORIES		$EU(a_i)$
	$A_1$	$A_2$	
$a_1: A_1$	1.0	0	$P(A_1)$
$a_2: A_2$	0	1.0	$P(A_2)$

then the expected utility for a given question is equivalent to the probability of the most probable alternative. Thus the optimal strategy is to choose the most probable alternative as the correct alternative. If both are equally probable, then an answer can be obtained by choosing each alternative with a probability of .5. Thus, the branching procedure in Figure 3 for the choice method is firmly based on decision theory (Shuford & Massengill, 1965).

Table 4 shows the alternative which will be chosen as the correct alternative for a given form of the question and a given category of

TABLE 4

The Alternative which will be Chosen as the Correct Alternative for a Given Form of the Question and a Given Category of Knowledge.

FORMS		CATEGORY		
		TRAINED	UNTRAINED	MISTRAINED
$F_1$	$A_1: A=B$	x	.5	
	$A_2: A=C$		.5	x
$F_2$	$A_1: A=C$		.5	x
	$A_2: A=B$	x	.5	

knowledge. If we compare Tables 2 and 4, we will see that whereas DM gives us unambiguous information about a subject's state of knowledge with one question, IM does not. For instance, if Form 1 of the question is asked and the student responds with  $A_1$  as the correct answer, then we know that he is not mistrained but we do not know, for certain, whether he is trained or untrained. If he responds with  $A_2$ , we know that he is



not trained, but we do not know, for certain, whether he is untrained or mistrained.

On the basis of this informal analysis, we can draw some conclusions about the two types of testing and the question of whether or not to test in this situation. At this point, we can say that DM guarantees correct classification of all students on the basis of one question, for our situation. If no test is used, correct classification of all students is only guaranteed at three points, i.e., where a given one of the categories has a probability of 1.0 of occurring. The use of IM improves on this somewhat by always giving correct classification not only when a given category has a probability of 1.0, but also when  $T$  and  $mT$  together have a probability of one. Thus, DM is better than either IM or classifying without testing (CWT), for most conditions, whereas the latter two are never better than DM. Also, the goodness of IM and CWT depends on the values of instructor's prior probabilities, whereas DM only depends on the data observed. This gives us a start toward our purpose of evaluating the two testing methods for the situation we have specified. But now we would like to know how much better, if any, DM is than IM for each possible condition. This will enable us to be more specific in our comparison of the two methods. In order to answer the question of how much better, we will need to get information for IM and DM which is analogous to the information for CWT summarized in Figure 2. This will require a formal analysis of the two methods in terms of the situation we defined.

#### 4. Formal Analysis of the Two Testing Methods.

In this formal analysis, we will only show the results of our derivations with comments on these results. To enable the reader who is interested to get some idea of the steps involved in the derivations, we include in Table 5 a summary of the basic decision-theoretic relationships used here.\*

TABLE 5

Summary of the Basic Decision-Theoretic Relationships to be used in the Derivations of this Report

Payoff matrix:

ACTS	CATEGORIES					
	$S_1$	$S_2$	$\dots$	$S_j$	$\dots$	$S_m$
$a_1$	$u_{11}$	$u_{12}$	$\dots$	$u_{1j}$	$\dots$	$u_{1m}$
$a_2$	$u_{21}$	$u_{22}$	$\dots$	$u_{2j}$	$\dots$	$u_{2m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_i$	$u_{i1}$	$u_{i2}$	$\dots$	$u_{ij}$	$\dots$	$u_{im}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_n$	$u_{n1}$	$u_{n2}$	$\dots$	$u_{nj}$	$\dots$	$u_{nm}$

Prior probabilities: probability of the category  $S_j$ ,

$$P(S_j); \text{ where } \sum_j P(S_j) = 1.0$$

Conditional probabilities: probability of data  $d_k$  given category  $S_j$ ,

$$P(d_k | S_j); \text{ where } \sum_k P(d_k | S_j) = 1.0$$

Unconditional probabilities: probability of data  $d_k$ ,

$$P(d_k) = \sum_j P(S_j) P(d_k | S_j); \text{ where } \sum_k P(d_k) = 1.0$$

---

\* See Raiffa and Schlaifer (1961) for the mathematical background leading to these relationships.

TABLE 5 (Cont.)

Posterior probabilities: probability of category  $S_j$  given data  $d_k$ ,

$$P(S_j | d_k) = \frac{P(S_j)P(d_k | S_j)}{P(d_k)} \quad \text{where } \sum_j P(S_j | d_k) = 1.0.$$

Expected utility of act  $a_i$  given data  $d_k$ :

$$\begin{aligned} EU(a_i | d_k) &= \sum_j P(S_j | d_k) u_{ij} \\ &= P(S_j | d_k); \text{ for the utility structure of Table E.1.} \end{aligned}$$

Maximum expected utility given data  $d_k$ :

$$EU(a^* | d_k) = \max_i EU(a_i | d_k)$$

Average expected utility of responding with the optimal act for each data result:

$$\begin{aligned} EU(a^*) &= \sum_k P(d_k) EU(a^* | d_k) \\ &= \sum_j P(S_j) P(d_k | S_j), \text{ for the utility structure of Table E.1,} \\ &= \text{Expected proportion of correct classifications.} \end{aligned}$$

#### 4.1 Direct Method.

At this point, we want to show formally that the average expected utility yielded by DM is equal to 1.0 for all conditions in the situation we are using. In other words, we want to show that DM will always lead to the correct classification of students in this situation. And in the process we will show that the instructor's prior probabilities are irrelevant.

Prior probabilities. There are three prior probabilities:  $P(T)$ ,  $P(uT)$ , and  $P(mT) = 1.0 - P(T) - P(uT)$ , i.e., the probability that a student is trained, the probability that he is untrained, and the probability that he is mistrained. The three probabilities sum to 1.0.

Data. For a given question, we can get one of three data results from the student. These are the probabilities 1.0, .5, and 0. As is evident from Table 2, the meaning of the data will depend on the form of the question used. Our derivations will be in terms of the first form. The results will be analogous for the second. Thus, the possible data results are:

$$d_1: P(A_1)=1.0,$$

$$d_2: P(A_1)=.5,$$

$$d_3: P(A_1)=0.$$

Conditional probabilities. We will now state formally the relevant conditional probabilities.

$$P(d_1|T) = 1.0.$$

$$P(d_2|uT) = 1.0.$$

$$P(d_3|mT) = 1.0.$$

Actually, we can talk about the probability of each data result given each category, but there is no need to do so in this case, since for a given category, one data result has all of the probability. Thus, if the student is trained, only  $d_1$  can occur; if he is untrained, only  $d_2$  can occur; and if he is mistrained, only  $d_3$  can occur.

Unconditional probability of the data. In this case, the probability of a given data result occurring is equal to the prior probability of the category which can yield that result. Thus,

$$\begin{aligned} P(d_1) &= P(T)P(d_1|T) + P(uT)P(d_1|uT) + P(mT)P(d_1|mT) \\ &= P(T), \end{aligned}$$

$$P(d_2) = P(uT),$$

$$P(d_3) = P(mT).$$

Thus, for example, the probability that  $d_1$  will occur is equal to the probability that the student is trained,  $P(T)$ .

Posterior probabilities. Now let us see how a particular data result affects the prior probabilities.

$$P(T|d_1) = P(T)P(d_1|T)/P(d_1) = 1.0$$

$$P(uT|d_2) = 1.0.$$

$$P(mT|d_3) = 1.0.$$

It is clear that a particular category is certain to occur once a given data result has been observed and that a different category is certain to occur for each data result. Thus, a data result implies, unambiguously, the state of knowledge of the student. As we can see from the equations, the prior probabilities have absolutely no effect on the posterior probabilities. One implication of this is that an instructor need have no prior knowledge of the student taking the test in order to classify him correctly in this situation.

Expected utility. Since for this situation the maximum expected utility for a given data result is equivalent to the posterior probability of the most probable category, only one expected utility is different from 0 for a given data result and that one is equal to 1.0. Thus, the maximum expected utilities are:

$$EU(a_1^*|d_1) = 1.0,$$

$$EU(a_2^*|d_2) = 1.0,$$

$$EU(a_3^*|d_3) = 1.0.$$



This means that the optimal acts are:  $a_1$ , if  $d_1$  is observed;  $a_2$ , if  $d_2$  is observed; and  $a_3$ , if  $d_3$  is observed. And from the equations we see that the expected utility of the optimal act for each possible data result is 1.0.

Average expected utility. Since an expected utility of 1.0 is guaranteed regardless of the data result obtained, the instructor is guaranteed an average expected utility of 1.0, i.e.,

$$EU(a^*) = 1.0.$$

This is true regardless of the values of the prior probabilities. Thus, over the whole surface shown in Figure 2, the expected utility for a one-item test is 1.0. This is an improvement over the approach of classifying without testing except at the three corners of the triangle. Of course, whether or not one should test with DM or classify without testing depends on the values of the prior probabilities and the cost of testing.

#### 4.2 Indirect Method.

We have seen that only one question is required for DM in order that each student be correctly classified, for the situation under discussion. We have also seen, informally, that there are situations in which IM does not allow for perfect classification, at least with one question. Thus, in the very beginning, we will include the idea of repeating the same question a number of times to see if this might improve the performance of one who uses IM. In deriving the results for IM, we have assumed that the student's answer to a question does not influence his answer to the same question when it is asked again. In other words, the machine always

behaves as if a question has not been previously encountered. (We will examine the implications of this assumption for IM in Section 5.1. For DM, of course, we do not have to worry about repeating items, at least in this situation, since one item is sufficient for perfect classification.)

Prior probabilities. These are the same as for DM.

Data. If a question is asked once, there are two possible pieces of information: the student is correct,  $C$ , or he is not correct,  $\sim C$ . From Table 4, we can see that if a student is trained, the data result will always be  $C$ . If he is mistrained, it will always be  $\sim C$ . But if he is untrained, it may be either. (It is this last possibility which brings ambiguity into the situation.)

If we ask the question  $n$  times, we will get  $r$   $C$ 's and  $n-r$   $\sim C$ 's. For the trained student, we can only get the result  $r=n$ , i.e.,  $n$  correct answers out of the  $n$  times the question is asked. We will denote this result as  $C_n$ . For the mistrained student, we can only get the data result  $r=0$ , i.e., no correct responses out of  $n$  questions. We will denote this result as  $C_0$ . For the untrained student, we can get either of these two results and also the result  $C_{r*}$ , where  $r$  may equal any integer between 1 and  $n-1$ . Thus, we are interested in three data results for CM:  $C_n$ ,  $C_0$ , and  $C_{r*}$ .

Conditional probabilities of the data. The following are the relevant conditional probabilities for any one trial:

$$P(C|T) = 1.0,$$

$$P(C|uT) = .5, P(\sim C|uT) = .5,$$

$$P(\sim C|mT) = 1.0.$$

Since the trials are independent, we can obtain the conditional probabilities for a given category of knowledge,  $S$ , by use of the binomial probability equation. Thus,

$$P(C_r|S) = \binom{n}{r} [P(C|S)]^r [P(\neg C|S)]^{n-r}.$$

As we have seen, only one data result,  $C_n$  has any probability for the trained student:

$$P(C_n|T) = 1.0.$$

Likewise, only one data result,  $C_o$  has any probability for the mistrained student:

$$P(C_o|mT) = 1.0.$$

For the untrained student, however, each of the three data results has some probability for a finite  $n$ :

$$P(C_n|uT) = \frac{1}{2^n}$$

$$P(C_o|uT) = \frac{1}{2^n}$$

$$\begin{aligned} P(C_{n*}|uT) &= 1 - [P(C_n|uT) + P(C_o|uT)] \\ &= \frac{2^{n-1}-1}{2^{n-1}} \end{aligned}$$

We can see immediately that as  $n$  approaches infinity, the last probability approaches 1.0 and the total situation tends to the one we had for DM, i.e., a particular data result implies a particular classification.

Unconditional probabilities. Now we want to look at the probability that a given data result will occur.

$$\begin{aligned} P(C_n) &= P(T) + \frac{1}{2^n} P(uT) \\ &\rightarrow P(T) \quad \text{as } n \rightarrow \infty. \end{aligned}$$

$$P(C_{n*}) = \frac{2^{n-1}-1}{2^{n-1}} P(uT) \\ \rightarrow P(uT) \quad \text{as } n \rightarrow \infty .$$

$$P(C_o) = \frac{1-2^n}{2^n} P(uT) + 1.0 - P(T) \\ \rightarrow P(mT) \quad \text{as } n \rightarrow \infty .$$

Here, again, as  $n$  approaches infinity, the values of these probabilities approach the same values they had for DM.

It should be noted that for  $n=1$ ,

$$P(C_{n*}) = 0,$$

since for  $n=1$   $r$  must equal 0 or 1. In other words only two of the data results are possible when  $n=1$ .

Posterior probabilities. Now we want to see how a particular data result affects the probabilities of the categories. First, let us see what happens when  $C_n$  is observed.

$$P(T|C_n) = \frac{P(T)}{P(T) + \frac{1}{2^n} P(uT)} \\ \rightarrow 1.0 \quad \text{as } n \rightarrow \infty .$$

$$P(uT|C_n) = \frac{\frac{1}{2^n} P(uT)}{P(T) + \frac{1}{2^n} P(uT)} \\ \rightarrow 0 \quad \text{as } n \rightarrow \infty .$$

$$P(mT|C_n) = 0.$$

Thus, as we have already observed, only  $T$  and  $uT$  have any probability when  $C_n$  is observed. And as  $n$  approaches infinity, only  $T$  has any.

When  $C_o$  is observed,

$$\begin{aligned}
P(T|C_0) &= 0, \\
P(uT|C_0) &= \frac{\frac{1}{2^n} P(uT)}{\frac{1-2^n}{2^n} P(uT) + 1.0 - P(T)} \\
&\rightarrow 0 \text{ as } n \rightarrow \infty, \\
P(mT|C_0) &= \frac{1-P(T) - P(uT)}{\frac{1-2^n}{2^n} P(uT) + 1.0 - P(T)} \\
&\rightarrow 1.0 \text{ as } n \rightarrow \infty.
\end{aligned}$$

Only  $uT$  and  $mT$  have any probability when  $C_0$  is observed and, as  $n$  approaches infinity, only  $mT$  has any.

When  $Cr^*$  is observed,

$$P(uT|Cr^*) = 1.0.$$

Thus, as  $n$  approaches infinity, the posterior probabilities take on the same values as they did for DM.

Optimal acts. For the utility structure of Table 1, the optimal act for a given data result depends on which of the posterior probabilities is largest. For the outcome  $C_n$ , the optimal strategy is to choose  $a_1$  when

$$P(T|C_n) > P(uT|C_n),$$

i.e., when

$$P(uT) < 2^n P(T) \equiv Y,$$

and to choose  $a_2$  when the inequality is reversed.

For the outcome  $C_0$ , the optimal strategy is to choose  $a_2$  when

$$P(uT|C_0) > P(mT|C_0),$$

i.e., when

$$P(uT) > \frac{2^n}{2^n + 1} - \frac{2^n}{2^n + 1} P(T) \equiv Z,$$

and to choose  $a_3$  when the inequality is reversed.

For the outcome  $C_{n^*}$  the optimal strategy is always to choose  $a_2$ .



Thus we see that the optimal act given  $C_n$  or  $C_o$  depends not only on the data observed but also on the relationship between  $P(uT)$  and  $P(T)$ . There are four different possible relationships between  $P(uT)$  and  $P(T)$ . These are shown as the row headings of Table 6. For a given one of these relationships, a given data result determines the optimal act. For example, when  $Z < P(uT) < Y$ ,  $a_1$  is the optimal act when  $C_n$  occurs. We can use these relationships to divide the area shown in Figure 1 into four sections:  $S_1$ ,  $S_2$ ,  $S_3$ , and  $S_4$ . Each of the four sections is characterized by one of the rows in Table 6, i.e., by a certain pattern for the optimal acts given the data.

Figure 4 shows the surface of possible prior probabilities divided into four sections as a function of the relationship between  $P(uT)$  and  $P(T)$ , for  $n=1$ . Each of the sections corresponds to one of the rows in Table 6. For example, the area labelled  $S_2$  corresponds to the second row in the table. The line in the figure labelled  $C_n$  represents the dividing line through the surface for the two possible optimal acts when  $C_n$  is

TABLE 6

SECTION	DATA		
	$C_n$	$C_{r*}$	$C_o$
$S_1: Z < P(uT) < Y$	$a_1$	$a_2$	$a_2$
$S_2: P(uT) < Y, Z$	$a_1$	$a_2$	$a_3$
$S_3: P(uT) > Y, Z$	$a_2$	$a_2$	$a_2$
$S_4: Z > P(uT) > Y$	$a_2$	$a_2$	$a_3$

$$Y = 2^n P(T); \quad Z = \frac{2^n}{2^n + 1} - \frac{2^n}{2^n + 1} P(T)$$

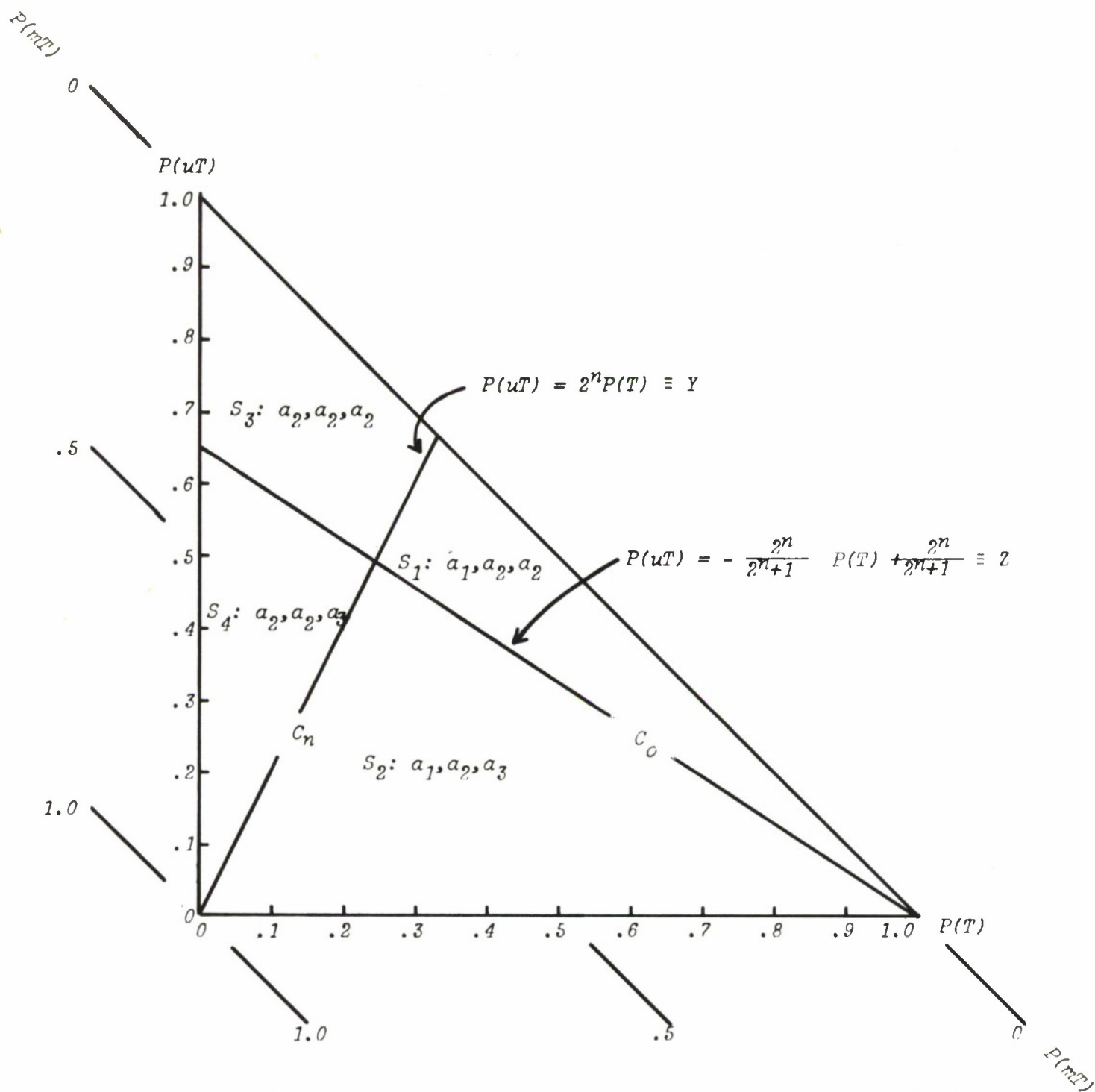


Figure 4. The division, for  $n=1$ , of the surface of possible prior probability combinations according to the pattern of optimal acts for the possible sample results. For any section, the acts listed are optimal given  $C_n$ ,  $C_{n*}$ , and  $C_0$ , respectively.

observed. For the points above that line

$$P(uT) > 2^n P(T),$$

i.e.,

$$P(T|C_n) < P(uT|C_n),$$

and, thus,  $a_2$  should be chosen. The inequality is reversed for the points below that line and so  $a_1$  should be chosen. Similarly, the line labelled  $C_0$  represents the dividing line for the case when  $C_0$  is observed. As we have seen, when  $C_{n*}$  is observed,  $a_2$  is always chosen regardless of the relationship between  $P(T)$  and  $P(uT)$ . Remember that these results are for  $n=1$ . We can divide the surface in analogous fashion for each possible value of  $n$ . Figure 5 shows the divisions ranging from  $n=1$  to  $n=4$ . Notice that as  $n$  gets larger,  $S_2$ , the section of the surface for which it is optimal to take  $a_1$ ,  $a_2$ ,  $a_3$  respectively, for the data results  $C_n$ ,  $C_{n*}$ ,  $C_0$ , gets larger. And notice that as  $n$  approaches infinity, the dashed line approaches the  $P(uT)$  axis and the solid line approaches the right hand boundary of the surface, i.e., there is only one optimal strategy for each data result regardless of the prior probabilities. Thus as  $n \rightarrow \infty$ , CM gives results for all conditions which are equivalent to those given by PM for one item.

Average expected utility. We can find the average expected utility for any point in any of the sections. To do this, we simply weight the expected utility of the optimal act given the data by the probability of the data, for each data result, and then sum over these weighted results. This gives us the average expected utility, given  $P(T)$  and  $P(uT)$ , of choosing the optimal act for each data result. Thus, for the four sections of our surface,

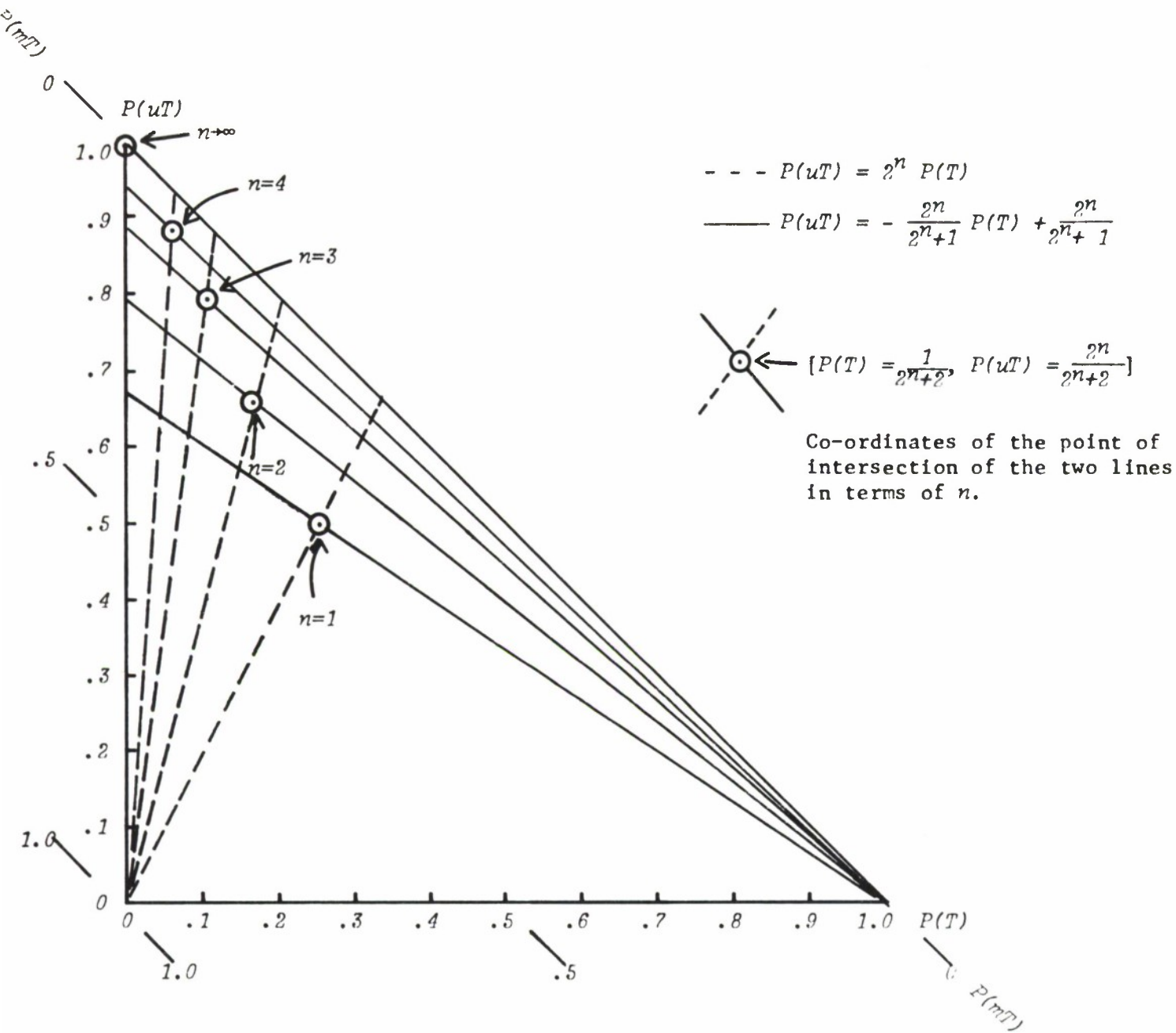


Figure 5. The changes in the four divisions of the probability combination surface as a function of  $n$ .

$$\begin{aligned}
EU(\alpha^*|S_1) &= \frac{2^n-1}{2^n} P(uT) + P(T) \\
&\rightarrow P(uT) + P(T) \text{ as } n \rightarrow \infty , \\
EU(\alpha^*|S_2) &= - \frac{1}{2^{n-1}} P(uT) + 1.0 \\
&\rightarrow 1.0 \text{ as } n \rightarrow \infty , \\
EU(\alpha^*|S_3) &= P(uT), \\
EU(\alpha^*|S_4) &= - \frac{1}{2^n} P(uT) + 1.0 - P(T) \\
&\rightarrow 1.0 - P(T) \text{ as } n \rightarrow \infty .
\end{aligned}$$

Figure 6 shows the average expected utility, for  $n=1$ , for selected points on the prior probability surface. The lowest  $EU$  in the figure is .5. This lowest value only occurs for one point (.25, .50, .25). As we go out from this point,  $EU$  gets larger. Compare these results with Figure 2, i.e., the case in which no test is given. There the worst possible  $EU$  is  $1/3$ . This occurs only for the point  $(1/3, 1/3, 1/3)$ . When no test is given, there are only three points, the three corners, for which  $EU(\alpha^*) = 1.0$ . These are the three possible cases for which two of the states have a probability of 0 while the remaining state has a probability of 1.0. But notice that when a one-item IM test is given, that besides these three points, there is a whole line, the  $P(T)$  axis, which gives an average expected utility of 1.0. These are the cases for which  $P(uT)=0$ , i.e., the student is either trained or mistrained. The reason the average expected utility is 1.0 for these cases is that the data discriminates perfectly when only  $T$  and  $mT$  are possible. In other words, a correct answer implies that the student is trained, while an incorrect answer implies that he is mistrained. Consult Table 4 to confirm this point.



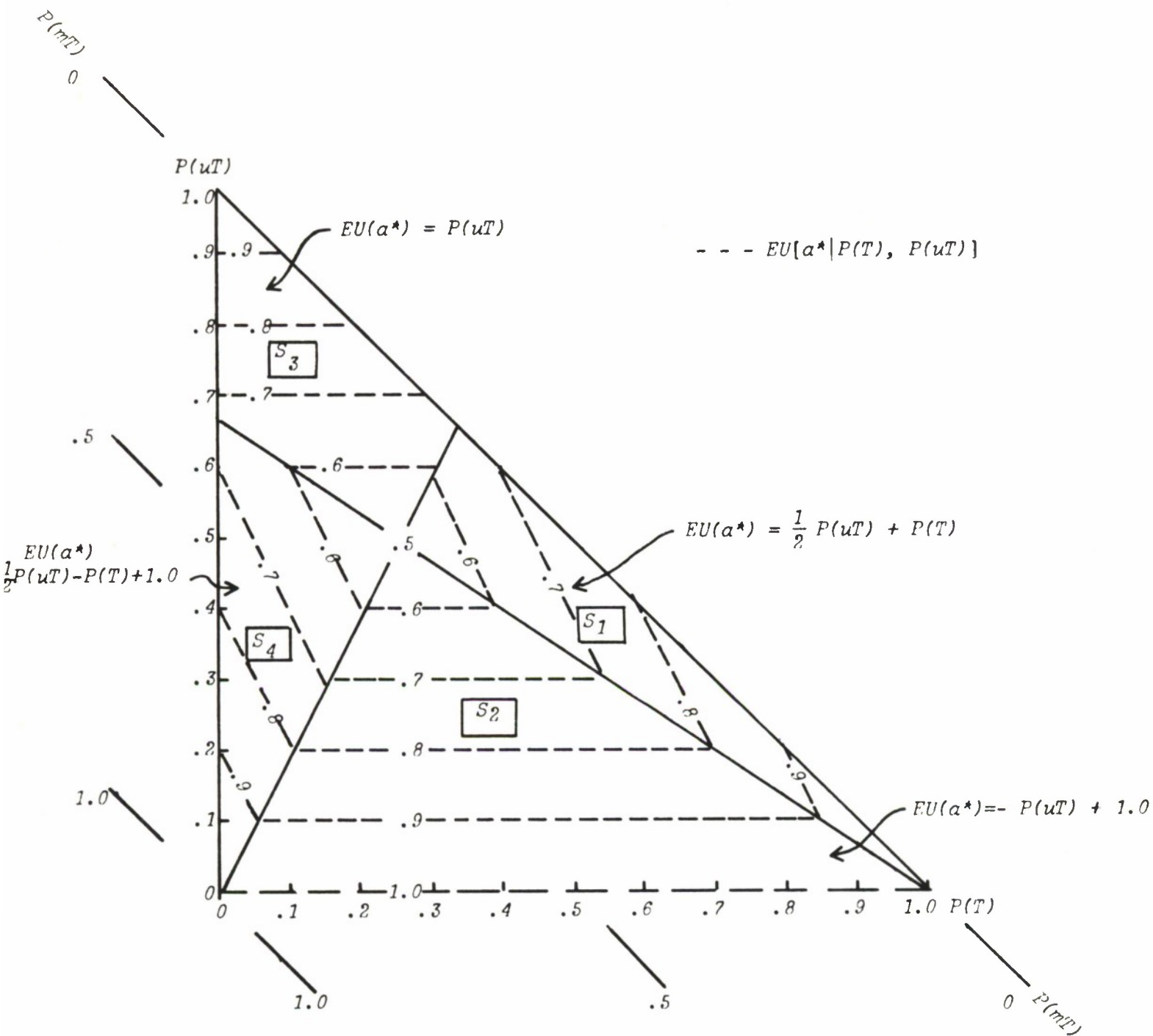


Figure 6. The average expected utility for selected points on the probability combination surface for  $n=1$ .

If we were to draw a figure analogous to Figure 6 for each value of  $n$ , we would see that the minimum average expected utility for a given  $n$  would be at the point where the two dividing lines cross. The coordinates of this point are

$$P(T) = \frac{1}{2^{n+2}} , P(UT) = \frac{1}{2^{n+2}}$$

and the average expected utility for this point is

$$\frac{2^n}{2^{n+2}} .$$

Table 7 shows the value of the minimum average expected utility for selected  $n$ 's. Note that as  $n$  approaches infinity, the minimum average expected utility approaches 1.0. Also note that even for eight questions, the minimum value is getting very large.

TABLE 7

The Minimum Average Expected Utility for the IM Test for Selected Values of  $n$ .

$n$	Minimum Average Expected Utility = $\frac{2^n}{2^{n+2}}$
0	.333...
1	.500
2	.666...
3	.800
4	.888...
5	.941
6	~.970
7	~.985
8	~.992
⋮	⋮
$\rightarrow \infty$	1.000

## 5. The Effectiveness of the Two Methods.

Before we discuss the effectiveness of the two methods, it should be reiterated that the conclusions which we draw are in terms of the specific response model we have assumed. Some or all of our conclusions may be valid in more complex situations, but that is a matter for further investigation.

To put our discussion of effectiveness into perspective, we will restate the main assumptions that have been made. It is assumed in the response model we are using that a student has probabilities for the alternatives of a question and that the values of these probabilities depend in a specific way on the state of his training. For the direct method of testing, it is assumed that these probabilities can be inferred by a measurement procedure. For the indirect method, it is assumed that the student uses them with an all-or-none payoff function to choose the alternative which will maximize his expected utility.

An all-or-none payoff function has also been assumed for the instructor who is making the classifications. This means that the expected utility of the instructor can also be interpreted as the expected proportion of correct classifications, and, of course, that maximization of expected utility can be interpreted as maximization of expected proportion of correct classifications.

The question of the effectiveness of the two methods can be looked at from two points of view. One is from the point of view of the instructor who is, at a given moment, classifying a student. The other

is from the point of view of some outside agent who knows the true state of each student and who can thus evaluate the instructor's performance in terms of the actual state of each student's knowledge. For both points of view, average expected utility is used as the measure of effectiveness.

#### 5.1 Effectiveness from the Instructor's Point of View.

First, we will look at effectiveness from the point of view of the instructor. Once the instructor has assigned prior probabilities\* to the three states for a particular student, and given that he accepts the response model and the payoff structure we have specified, the results of our derivations furnish him with an expected utility for each act given each possible data result as well as average expected utilities for responding with the optimal act for each data result. Thus he can evaluate the effectiveness of each method in terms of its average expected utility,  $EU(a^*)$ .

Having related the results of our derivations with effectiveness, from the instructor's point of view, let us briefly review CWT, IM, and DM in terms of  $EU$ . For classification without testing, CWT,  $EU$  ranges from .33... to 1.0. For IM, the range of  $EU$  depends on  $n$ . Table 7 shows the lower bound of the range for various values of  $n$ . The upper bound for IM is 1.0 regardless of  $n$ . For DM,  $EU$  is a constant, 1.0. Thus, DM has the narrowest range of  $EU$ . And aside from the cost of using the methods, DM is better than or equal to either CWT or IM for all

---

\*

These probabilities are prior to the results of the testing but may include various types of non-test information about the student.

conditions. Of course, a person choosing between the methods would take cost into account. We do not do it here because it is not relevant to our arguments but our results are in a form which will enable anyone who is interested to do so.

Classifying without testing gives an *EU* of 1.0 at three points; the points for which a particular state is given a probability of 1.0. IM gives an *EU* of 1.0 at these points as well as at all of the points where  $P(uT)=0$ . And, of course, DM gives an *EU* of 1.0 at all points. Note that the points for which CWT and IM give 1.0 all require some form of certainty, either that a particular category is the case or that a particular category is not the case.

Influence of instructor's prior on effectiveness. For CWT, the closer a prior is to one of the three corners, the larger *EU* is (See Figure 2). For IM, the closer a prior is toward the corner for which  $P(uT)=1.0$  or toward the line for which  $P(uT)=0$ , the larger *EU* is. This means that for CWT and IM, the instructor may be able to use background information on a particular student in conjunction with his observations on that student during the lesson to increase his *EU* for the student.

By observing students during a lesson and by connecting his observations with background information on the students, the instructor may get some idea of what percentage of the group will fall in each category. He could use this information to obtain a single prior which would be used for each student. His effectiveness in classification might be very good but there is room for objection to his use of a single prior since it is, in effect, using a group average to classify individual students. He could remedy



this situation by recalling what he had observed about particular students and attempting to assign a prior to each student reflecting his feelings about the state of knowledge of that particular student. And further, he could use his feelings concerning the group as a whole to check the coherence (de Finetti, 1937) of his priors for individuals.

Thus the instructor may be able to improve the effectiveness of CWT and IM by obtaining relevant background information on his students and by observing them during the lesson. Certainly, this is an improvement over approaches which use only part of the available information to classify students and which use that information to classify a student not in terms of his absolute performance but in terms of the performance of some group of which he is a member.

Since relevant information about individual students is essential to CWT and IM, but not to DM, it is easy to see the contribution that DM can make in situations, conforming to our assumptions, in which the person making the classifications may not be on hand to observe the student, e.g., self-instruction, instruction by television; or in which there are large numbers of students in a class thereby handicapping the instructor in obtaining information about individual students. But regardless of how much information the instructor is able to obtain about his students, his performance with CWT and IM will never be better than with DM, for the situation we are considering.

Our comments on the instructor's prior point up the fact that there is information other than answers to test items which can be taken into

account in classifying students. To the extent that this information can increase the instructor's certainty about the state of a student, CWT and IM increase in effectiveness. But for the situation we have defined, CWT and IM are never more effective than DM. Thus, if cost, in conjunction with effectiveness, justifies the use of DM, we can skirt the whole issue of the instructor's probabilities, since the information incorporated in them is superfluous. \*

The reader should be clear on the reason that IM is less effective for most conditions than DM. The reason does not lie in the area of the instructor's subjective probabilities. The derivation of both IM and DM involved the instructor's subjective probabilities. The difference is in the conditional probabilities yielded by IM as opposed to those yielded by DM. The conditional probabilities of DM simply supply more information than those of IM. Thus, the fact that IM is less effective than DM cannot be taken as a deprecation of subjective probabilities. And, of course, the adoption of DM would not eliminate subjective probabilities from our consideration since the student's subjective probabilities are basic to the direct method.

As  $n$  gets larger, the prior probabilities of the instructor become less important in the case of most priors and the effectiveness of IM approaches that of DM. And, as we have seen, the approach of IM to DM in terms of performance is quite rapid so that  $n$  does not have to be very large for IM to approximate DM (See Table 7). This brings us to the question of independence of trials.

---

\* It should be noted that by eliminating the need for the instructor's prior probabilities and thus allowing a larger class to be taught with no loss in effectiveness this economic benefit of DM should certainly affect the slight additional cost of testing with DM.

Independence of trials. We have assumed for IM that the test items for a given concept are regarded by the student as being independent. This assumption would seem to put an extreme restriction on the possible applications of our results for IM. It is difficult to picture real-life situations in which we can be sure that the answer to a question will not affect a subsequent answer to the same question especially if it is asked again immediately.

Since, in our model, students who are either trained or mistrained would always give the same answer to repetitions of the question as they gave the first time it was asked while students who are untrained would not, it is the untrained student for whom the independence assumption is critical. Suppose, for example, that an untrained student followed the strategy, which could be optimal in terms of his formulation of the task, of giving the same answer to a particular question each time it is repeated that he gave the first time it was asked. This means that only the first trial would have any value in classifying the student and that the results for the  $n$  corresponding to the number of times the question was asked would be misleading. Thus, the results we have derived for IM, for  $n > 1$ , apply only if the trials are independent.

This means that IM is restricted to situations for which the trials are independent or that additional assumptions must be made in order to handle the case for  $n > 1$ . But DM is applicable without further assumptions regardless of whether the trials are independent.

Summary. Now let us summarize our conclusions regarding the effectiveness of the two methods from the standpoint of the instructor. We will do

so in terms of three values of  $n$ :  $n=0$ ,  $n=1$ , and  $n \rightarrow \infty$ . Remember that we are not taking into account the cost of using the methods.

For  $n=0$ , DM is at least as good as classifying without testing. The two procedures are equivalent only when the instructor is certain that a student is in a particular one of the three categories. In cases where certainty is lacking and the instructor has little relevant non-test information on the student, DM does much better than CWT.

For  $n=1$ , DM is at least as good as IM for all conditions. IM is equivalent to DM only when there is certainty that the student is untrained or when there is certainty that he is not untrained. And here again, since the prior probabilities are important for IM, DM will do much better than IM when certainty is lacking and the instructor has little relevant non-test information on the student.

As  $n$  gets large, the role of the prior probabilities lessens and the effectiveness of IM increases. As  $n \rightarrow \infty$ , IM approaches DM in effectiveness. But if more than one question is used for IM, the trials must be independent or more assumptions must be made in order for the results of IM to be meaningful. Of course, this necessity for independence does not apply to DM, since only one question is necessary in order to give perfect classification. Thus, if the instructor does not know whether the independence assumption applies in a situation and he does not have enough non-test information to tell him for certain which state a given student is in, then DM will outperform IM. We should also note that for IM test situations in which a very large number of questions are asked, the cost of using IM will finally come into play even if it is negligible for small  $n$ 's.

## 5.2 Effectiveness from an Outside Agent's Point of View.

We have discussed the effectiveness of the two test methods from the point of view of the instructor who is classifying students. Now we want to look at the same question from the point of view of an outside agent who knows the actual state of each student at the time the student is classified. The outside agent is in a position to evaluate an instructor, and thus to evaluate  $IM$  given the instructor, in terms of information in addition to that which the instructor has\*. We might point out that for our purpose it does not matter whether there is an agent who actually possesses a knowledge of the category of each student, since the conclusions we draw will be the same whether or not anyone actually has this knowledge.

The first step in the agent's procedure is to classify students who have already been classified by the instructor. Whereas the instructor classified on the basis of  $T$ ,  $uT$ , and  $mT$ , the agent classifies on the basis of the particular prior distribution the instructor used for a given student. We will represent an instructor's prior distribution by  $P$ , where  $P$  is the vector  $[P(T), P(uT), P(mT)]$ . As we have seen, Figure 1 shows all of the possible priors. Once the agent has classified students in terms of  $P$ , he can find the relative frequency with which the students, for whom a particular  $P$  was used, actually fell in  $T$ ,  $uT$ , and  $mT$ . We will designate this relative frequency distribution by  $F$ , where  $F$  is the vector  $[F(T), F(uT), F(mT)]$  and where  $F(T)$ ,  $F(uT)$ ,  $F(mT)$  designate the proportion

---

\*

Note that the agent need only be concerned with  $IM$  not  $DM$  since  $DM$  is independent of the instructor and guarantees an  $EU$  of 1.0 for all conditions.



of students, classified by the instructor, who are actually trained, untrained, and mistrained, respectively.

Now the agent is in a position to ask the following question: "What would the instructor's average expected utility be if the students for whom he uses  $P$  are actually distributed according to  $F$ ?" We will designate this average expected utility as  $EU(P|F)$ . Now let us see how we can obtain this average expected utility.

We have seen, in the case of IM, that the optimal pattern of acts for the possible data results depends on the prior distribution used by the instructor. (See Table 6). According to the results given in this table, one of four distinct patterns of action is optimal for each possible prior, i.e., for each  $P$ . Thus, for a particular  $P$ , an instructor can use the results of Table 6 to determine the pattern of acts, given the data, which will maximize his average expected utility. If the instructor gives the pattern of acts associated with  $P$ , when  $F$  is the relative frequency distribution of the actual states of the students for whom  $P$  is used, then the instructor would be expected to obtain  $EU(P|F)$  per student rather than  $EU(a^*)$ . Thus, at any point, the agent has an index,  $EUF = EU(P|F)$ , of the instructor's performance, so far.

It may be helpful at this point to distinguish between  $EUF$  and the actual proportion of correct classifications that the instructor has made for a given  $P$  at the time the agent is evaluating his performance. The actual proportion of correct classifications depends, at any point, on the distribution of data results generated by the students who are actually untrained. We have seen that in the long run this distribution

will be a function of a binomial distribution. But the data results will not, in general, be generated systematically. In other words, they will not have the form of our theoretical distribution at every point. For example, there may be a run of  $C_n$ 's so that at a given point many more  $C_n$ 's have been given by untrained students than our equations would indicate. Of course, in the long run, the data results generated by untrained students should approach the values given by our equations.

Thus the actual number of correct classifications an instructor has made, up to a given point, may not reflect how well he is using the information available to him. In other words, chance fluctuations in the data results may make it appear that he is using the available information better or worse than he actually is. To get rid of this effect, we use the theoretical values of  $P(d|S)$  rather than the actual percentages of  $d$  given  $S$  when computing  $EUF$ . Thus,  $EUF$  gives the amount the instructor would have made per student, by using  $P$  when  $F$  is the case, if the data results had been generated according to our equations up to this point. And so, arbitrary fluctuations of the data results do not affect the agent's evaluation of an instructor at a given point.

We have said that each prior can be associated with one of four distinct patterns of action. Since there are only four possible patterns of action, the agent needs only four graphs, for a particular value of  $n$ , in order to be able to obtain  $EUF$  for any  $P$  and  $F$ . This is because the equation for  $EU(P\epsilon S_i|F)$  is identical to the  $EU$  equation we have already derived for  $P\epsilon S_i$  when  $F$  is substituted for  $P$ . And further, the equation

for  $EU$  applies over the whole  $P$  surface for any  $P \in S_i$ . The four relevant equations are:

$$EU(P \in S_1 | F) = \frac{2^{n-1}}{2^n} F(uT) + F(T) ,$$

$$EU(P \in S_2 | F) = \frac{1}{2^{n-1}} F(uT) + 1.0 ,$$

$$EU(P \in S_3 | F) = F(uT) ,$$

$$EU(P \in S_4 | F) = - \frac{1}{2^n} F(uT) + 1.0 - F(T) .$$

Using these four equations, we can construct the four graphs for any  $n$ . Figure 7 shows the graphs for  $n=1$ . Notice that  $P$  determines which of the four graphs is relevant for a particular situation. For example, if  $P \in S_3$ , the agent would refer to the upper left hand graph. Once the graph is chosen, the relevant point on the graph is found by taking the point corresponding to  $F$ . Also notice that the range of  $EU$  is from 0 to 1.0 for each of the graphs. In other words, when the instructor gives  $P$  for  $X$  students and the frequency distribution of the actual states of the  $X$  students is  $F$ , the average expected utility,  $EU$ , which he could have been expected to make in this situation, could be anything between 0 and 1.0 depending on  $P$  and  $F$ .

To clarify the agent's procedure of evaluation, let us look at an example. Table 3 shows eight classifications by an instructor and the trial-by-trial evaluation, by the agent, of the instructor and thus of IM given the instructor. For the first subject  $P = (1.0, 0, 0)$ . (This prior falls on the border line between two sections. It will be sufficient for the comparisons we are going to make to regard it as being in  $S_2$ .) Thus the instructor would use the pattern  $a_1, a_2, a_3$  for the data results  $C_n, C_{n*}$  and  $C_o$ , respectively. The average expected

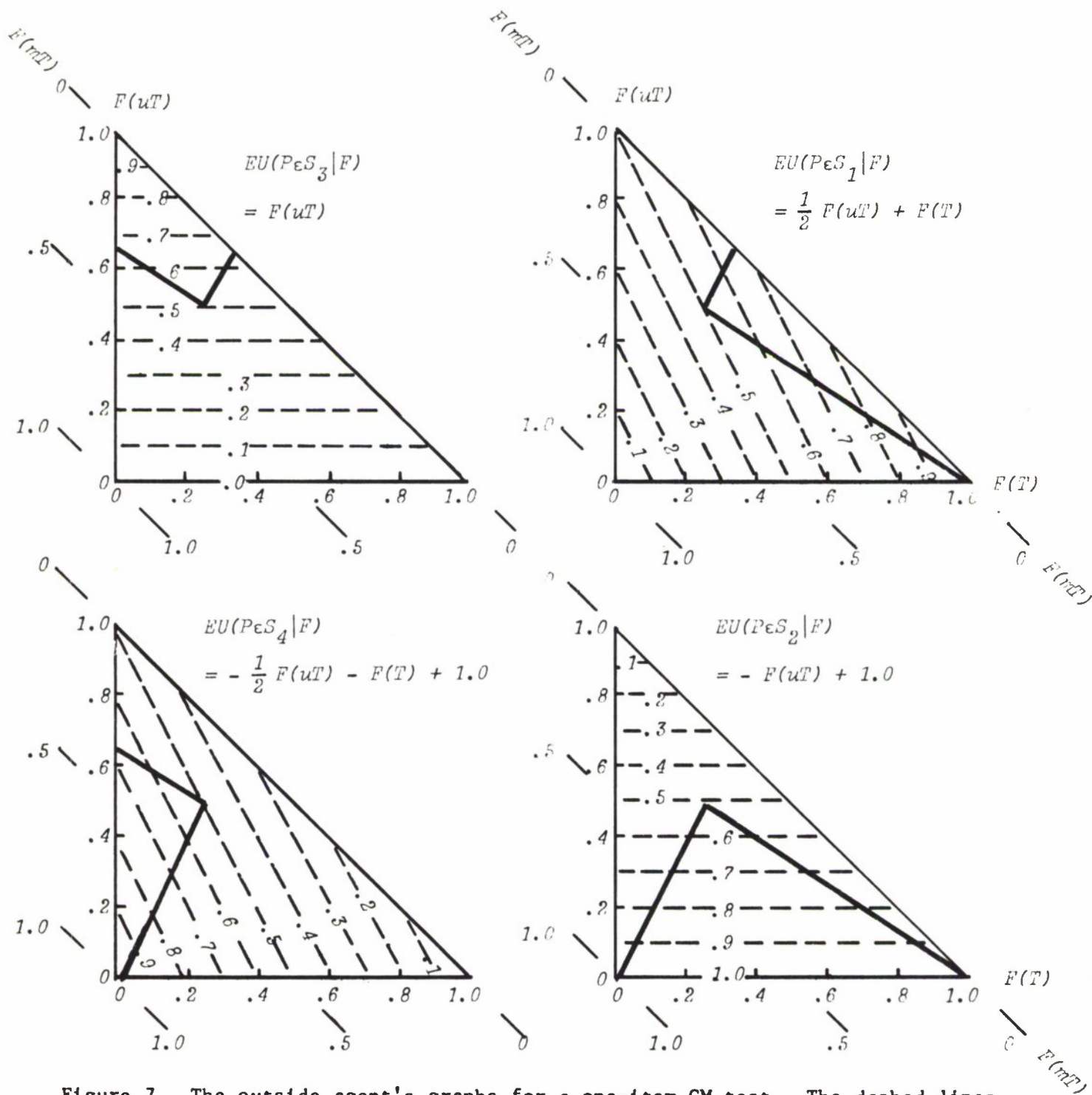


Figure 7. The outside agent's graphs for a one-item CM test. The dashed lines show the average expected utility of using a  $PeS_i$  given that the actual states of the students for whom  $P$  is used are distributed as  $F$ .

TABLE 8

Example of trial-by-trial evaluation of instructor by the outside agent where a trial is the classification of one student by the instructor in terms of a one-item ( $n=1$ ) CM test.

Trial	Instructor's Prior			Section in which Prior located for $n=1$	Actual Category of Student	Current relative frequency distribution for instructor's prior			EU	EUF
	$P(T)$	$P(uT)$	$P(mT)$			$F(T)$	$F(uT)$	$F(mT)$		
1.	1.0	.0	.0	$S_2$	$T$	1.00	.00	.00	1.00	1.00
2.	.9	.1	.0	$S_1$	$T$	1.00	.00	.00	.95	1.00
3.	1.0	.0	.0	$S_2$	$uT$	.50	.50	.00	1.00	.50
4.	.3	.4	.3	$S_2$	$uT$	.00	1.00	.00	.60	1.00
5.	.9	.1	.0	$S_1$	$uT$	.50	.50	.00	.95	.75
6.	1.0	.0	.0	$S_2$	$T$	.66	.33	.00	1.00	.66
7.	.8	.1	.1	$S_2$	$T$	1.00	.00	.00	.90	1.00
8.	1.0	.0	.0	$S_2$	$mT$	.50	.25	.25	1.00	.75



utility,  $EU$ , for the instructor is 1.0. This is given in the next to last column. The actual state of the student is  $T$ . Thus, the current relative frequency distribution for  $P = (1.0, 0, 0)$  is  $F = (1.0, 0, 0)$ . And the average expected utility,  $EU_F$ , of using the pattern given by  $P$  when  $F$  is the case is 1.0.

For the second student, the instructor uses a different prior. Note that  $EU$  is less than  $EU'$  here. In other words, if all of the students for whom the instructor used this prior were trained, the instructor would classify them all correctly in the long run in spite of the fact that his  $EU$  is merely .95. This is because if all of the students were trained, only the data result  $C_n$  would be generated and, with this prior, the optimal strategy is to call the student trained when  $C_n$  is observed.

On trial three, the instructor uses  $P = (1.0, 0, 0)$  again. But this time the student is actually untrained. For  $n=1$  an untrained student can be either  $C_n$  or  $C_o$ . This means that there is a possibility of conflict between the instructor's prior and the data result, since the instructor has expressed certainty that the student is trained. If  $C_n$  is obtained from the untrained student, then the instructor will not be aware of the conflict. His  $EU$  will be 1.0. But since this student is untrained, the instructor will be unable to correctly classify all students for this prior. If  $C_o$  is obtained the instructor will either have to re-evaluate his prior or ignore the data. If the instructor in Table 8 obtained a  $C_n$  for trial 3, or if he obtained a  $C_o$  and ignored it, his would be .5.\*

---

\* Our comments on the first three trials can be used as an aid in examining the remaining trials in the table.

Table 9 shows the summary measures for Table 8. The instructor has used  $P = (1.0, 0, 0)$  four times and the actual states of the students have

TABLE 9

Summary measures of Table 8 showing instructor's performance to date from viewpoint of outside agent.

Instructor's Prior			Section	Current relative frequency distribution for instructor's prior			EU	EUP
$P(T)$	$P(uT)$	$P(mT)$		$F(T)$	$F(uT)$	$F(mT)$		
1.0	.0	.0	$S_2$	.50	.25	.25	1.00	.75
.3	.1	.0	$S_1$	.50	.50	.00	.95	.75
.3	.4	.3	$S_2$	.00	1.00	.00	.60	1.00
.8	.1	.1	$S_1$	1.00	.00	.00	.90	1.00

been distributed as  $F = (.50, .25, .25)$ . Thus, as far as the instructor is concerned, his average EU for the four trials is 1.0. But from the standpoint of the agent, it is .75. This points up a difference between PM and CM. If the instructor had used DM, he would have been guaranteed the correct classification of all four students. But with IM, he is not guaranteed the correct classification of each student, even though  $P = (1.0, 0, 0)$  and  $EU = 1.0$ .

Thus, we see that IM involves more uncertainty than DM. And the additional uncertainty in IM comes from the fact that IM is dependent on the prior probabilities of the instructor whereas DM is not. If it were known for certain that an instructor's  $P$  was equivalent to  $P'$ , then

then there would be no more uncertainty concerning  $IM'$ , for that instructor, than there is for  $DM$ . And,  $EU$  could be interpreted as both the instructor's average expected utility for that trial and the earnings per trial or the proportion of correct classifications per trial which could be expected in the long run. In other words,  $EU$  and  $EUF$  would be equivalent. Under these circumstances, we could say that there are certain conditions for which  $IM$  and  $DM$  are equivalent, namely, the conditions for which the instructor is certain that the student is untrained or certain that he is not untrained. And, of course,  $CWT$  would be equivalent to  $DM$  for the cases in which the instructor gives a prior probability of 1.0 to a particular category. Of course, these equivalences are from the agent's point of view.

But if  $P$  and  $F$  are not equivalent for an instructor, then  $EU$  and  $EUF$  will not, in general, be equivalent. Thus, we cannot, without making further assumptions, say what level of effectiveness we can expect from  $IM$  for a given instructor. But we do know that it can be no greater than 1.0 regardless of the relation of  $P$  to  $F$ .\*

### 5.3 Summary

It seems clear, after having compared the effectiveness of the two methods both from the instructor's point of view and from an outside agent's point of view, that the direct method is more effective than the indirect method for all conditions, aside from the question of the cost of use. From the instructor's point of view, there are conditions for which the two methods give equivalent results. But from the agent's point of view

---

\*

The questions raised in this section concerning the subjective probabilities of the instructor are analogous to questions which will become relevant in terms of the student's subjective probabilities when we begin to look at situations in which the student's subjective probabilities can be values anywhere in the interval  $[0, 1.0]$ .

we see that there is uncertainty involved in IM which is not involved in DM, viz., we are never certain that the instructor's  $P$  and  $F$  are related in such a way that  $EU$  and  $EU_F$  are both 1.0.

We have seen also that the effectiveness of IM can be improved up to a limit of  $EU=1$  if the instructor has relevant non-test information on the student and/or if a question is repeated. But regardless of the amount of additional information, the effectiveness of IM can never be greater than that of DM, since  $EU$  for DM is 1.0 for all conditions. We also noted that repeated questions are valid for IM, in our situation, only if the questions are treated by the student as being independent.

## REFERENCES

- Finetti, Bruno de (1937) La prèvision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincarè, 7. [Translated and reprinted as "Foresight: its logical laws, its subjective sources." in Kyburg, Henry E., Jr. & Smokler, Howard E. (Eds.) Studies in subjective probabilities. New York: Wiley, 1964.]
- Massengill, H. Edward (1964) Purposive systems: theory and application. ESD-TDR-64-531, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.
- Raiffa, Howard & Schlaifer, Robert (1961) Applied statistical decision theory. Boston: Division of Research, Harvard Business School.
- Shuford, Emir H., Jr. (1965) Cybernetic testing. ESD-TR-65-467, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.
- Shuford, Emir H., Jr., Albert, Arthur, & Massengill, H. Edward (1965) Admissible probability measurement procedures. ESD-TR-65-567, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.
- Shuford, Emir H., Jr. & Massengill, H. Edward (1965) On communication and control in the educational process. ESD-TR-65-568, Decision Sciences Laboratory, L. G. Hanscom Field, Bedford, Mass.



## DOCUMENT CONTROL DATA - R&amp;D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Decision Sciences Laboratory Electronic Systems Division L. G. Hanscom Field, Bedford, Mass.		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP N/A	
3. REPORT TITLE DIRECT VS INDIRECT ASSESSMENT OF SIMPLE KNOWLEDGE STRUCTURES			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) None			
5. AUTHOR(S) (Last name, first name, initial) Massengill, H. Edward Shuford, Emir H., Jr.			
6. REPORT DATE March 1966		7a. TOTAL NO. OF PAGES 51	7b. NO. OF REFS 6
8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S) ESD-TR-65-542	
b. PROJECT NO. 2806			
c. TASK 280609		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None	
d.			
10. AVAILABILITY/LIMITATION NOTICES Distribution of this document is unlimited.			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, USAF, L. G. Hanscom Field, Bedford, Mass.	
13. ABSTRACT <p>This report compares two types of classroom testing in terms of efficacy in guiding instruction. One type of testing is the traditional indirect method based on the observation of choices. The other type is the direct method based on admissible probability measurement. The general finding is that the direct methods always perform as well as and in most cases better than the indirect methods. This deficiency in the indirect method can be alleviated in theory by introducing redundancy into the test and asking the same question over and over again. The performance of indirect methods depends in a very critical manner upon the information available to the instructor from other sources about the current state of knowledge of each student. The performance of the direct methods is unaffected by this. The gain in effectiveness achieved by using direct methods must be balanced off against the cost of using these new methods. A direct method may require more student time per item than does an indirect method. This, however, may be more than compensated for by the requirement for redundancy when using the indirect method. In addition, since a direct method does not require additional information from the instructor as to the current state of knowledge of each student, the possibility exists that much larger classes may be taught with no loss in effectiveness thus implying even further economic benefits from the use of direct methods to guide classroom instruction.</p>			

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Classroom Testing Methods Cost Effectiveness of Testing Methods Guided Instructions						

## INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those

imposed by security classification, using standard statements such as:

- (1) "Qualified requesters may obtain copies of this report from DDC."
- (2) "Foreign announcement and dissemination of this report by DDC is not authorized."
- (3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through \_\_\_\_\_."
- (4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through \_\_\_\_\_."
- (5) "All distribution of this report is controlled. Qualified DDC users shall request through \_\_\_\_\_."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.